PRIFYSGOL ABERYSTWYTH UNIVERSITY

Aberystwyth University

Distance Measure Assisted Rough Set Feature Selection

Shen, Qiang; MacParthaláin, Neil Seosamh; Jensen, Richard

DOI:
10.1109/FUZZY.2007.4295518

Publication date:
2007

Citation for published version (APA):
Shen, Q., MacParthaláin, N. S., & Jensen, R. (2007). *Distance Measure Assisted Rough Set Feature Selection*. 1084-1089. https://doi.org/10.1109/FUZZY.2007.4295518

General rights
Copyright and moral rights for the publications made accessible in the Aberystwyth Research Portal (the Institutional Repository) are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the Aberystwyth Research Portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the Aberystwyth Research Portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

tel: +44 1970 62 2400
email: is@aber.ac.uk

Download date: 04. May. 2023

Distance Measure Assisted Rough Set Feature Selection

Neil Mac Parthaláin, Qiang Shen, and Richard Jensen

Abstract— **Feature Selection (FS) is a technique for dimensionality reduction. Its aims are to select a subset of the original features of a dataset which are rich in the most useful information. The benefits include improved data visualisation, transparency, a reduction in training and utilisation times and potentially, improved prediction performance. Many approaches based on rough set theory have employed the dependency function which is based on the information contained in the lower approximation as an evaluation step in the FS process with much success. This paper presents a novel rough set FS technique which uses the information of both the lower approximation dependency value and a distance metric for the consideration of objects in the boundary region. The use of this measure in rough set feature selection can result in smaller subset sizes than those obtained using the dependency function alone.**

I. Introduction

The principal aim of feature selection is to choose a subset of the original features present in a given dataset which provide the most useful information. Following selection, the most important information of the dataset should still remain. In fact, efficient FS techniques should be able to detect and ignore noisy and misleading features. As a result, the dataset quality may even *increase* through feature selection.

Classifier accuracy can be increased as a result of feature selection, through the removal of noisy or misleading features. Also in domains where features correspond to measurements (the water treatment plant in [11] demonstrates this well), fewer features obviously offer advantages such as mimimising the expense and time consumed taking such measurements. For those datasets, which are smaller in size the runtimes of learning algorithms can be improved significantly. This is equally applicable to both training and application (e.g. classification) phases. Reduction of the data to fewer dimensions, also leads to the easy identification of trends within the data. This becomes evident where few features have an influence on data outcomes.

Methods which extract knowledge from data (e.g. rule induction) may also benefit from the use of FS and show improvement in the readability of the discovered knowledge. When induction algorithms are applied to reduced data, the resulting rules are more compact. A good feature selection process will remove unnecessary attributes which may affect both rule comprehension and rule prediction performance.

The work on rough set theory (RST) [9] offers a formal methodology that can be employed to reduce the dimensionality of datasets, as a preprocessing step to assist any chosen modelling method for learning from data. It assists in selecting the most information-rich features in a dataset. This is achieved without transforming the data, whilst simultaneously attempting to minimise information loss during the selection process. In terms of computational effort, this approach is highly efficient, and based on simple set operations, which makes it suitable as a preprocessor for techniques that are much more complex. In contrast to statistical correlation-reduction approaches [4], RST requires no human input or domain knowledge other than the given datasets. Perhaps most importantly though, it retains the underlying semantics of the data, which results in models that are more transparent to human scrutiny.

Most existing rough set-based FS approaches such as rough-set attribute reduction (RSAR) [2] rely on the information gathered from the lower approximation of a set to minimise data. These approaches although successful, ignore the information that is contained in the boundary region, or region of uncertainty. Whilst there are also some existing RST approaches which consider the boundary region information [3], [5], they adopt an approach which examines the upper approximation as a whole rather than examining the lower approximation and boundary region as conceptually separate entities. This paper presents a method which examines both the information in the lower approximation and the information contained in the boundary region for the selection of feature subsets.

The remainder of this paper is structured as follows. Section 2 summarises the theoretical basis and ideas of RSAR, along with a look at the rough set QUICKREDUCT algorithm. Section 3 describes a distance-metric assisted approach to RSAR (DMRSAR) and corresponding algorithm. Section 4 shows the results of applying fuzzy-rough feature selection FRFS [6], and DMRSAR approaches to a number of datasets, along with a comparison of run times, classification accuracies (using a fuzzy classifier), and dimensionality reduction. Section 5 concludes the paper along with some suggestions, as well as a discussion of future work.

II. Rough Set Attribute Reduction

The principal focus of this paper lies in distance metric-assisted rough set attribute reduction (DMRSAR), however an in-depth view of the current RSAR methodology is necessary to appreciate the DMRSAR approach fully.

At the heart of the RSAR approach is the concept of indiscernibility. Let $I = (U, A)$ be an information system, where U is a non-empty set of finite objects (the universe) and A is a non-empty finite set of attributes so that $a : U \rightarrow V_a$ for every $a \in A$. V_a is the set of values that a can take. For

Neil Mac Parthaláin (email: nsm03@aber.ac.uk), Qiang Shen (email: qqs@aber.ac.uk), and Richard Jensen (email: rkj@aber.ac.uk), are with the Department of Computer Science, University of Wales, Aberystwyth, Wales, UK.

1-4244-1210-2/07/$25.00 ©2007 IEEE.

any $P \subseteq A$, there exists an associated equivalence relation $IND(P)$:

$$IND(P) = \{(x, y) \in U^2 | \forall a \in P, a(x) = a(y)\} \quad (1)$$

The partition generated by $IND(P)$ is denoted $U/IND(P)$ and is calculated as follows:

$$U/IND(P) = \otimes\{a \in P : U/IND(\{a\})\} \quad (2)$$

where,

$$A \otimes B = \{X \cap Y : \forall X \in A, \forall Y \in B, X \cap Y \neq \emptyset\} \quad (3)$$

If $(x, y) \in IND(P)$, then x and y are indiscernible by attributes from P. The equivalence classes of the P-indiscernibility relation are denoted $[x]_p$. Let $X \subseteq U$. X can be approximated using only the information contained in P by constructing the P-$lower$ and P-$upper$ approximations of X:

$$\underline{P}X = \{x | [x]_p \subseteq X\} \quad (4)$$

$$\overline{P}X = \{x | [x]_p \cap X \neq \emptyset\} \quad (5)$$

Let P and Q be equivalence relations over U, then the positive, negative and boundary regions can be defined:

$$POS_P(Q) = \bigcup_{X \in U/Q} \underline{P}X \quad (6)$$

$$NEG_P(Q) = U - \bigcup_{X \in U/Q} \overline{P}X \quad (7)$$

$$BND_P(Q) = \bigcup_{X \in U/Q} \overline{P}X - \bigcup_{X \in U/Q} \underline{P}X \quad (8)$$

By employing this definition of the positive region it is possible to calculate the rough set degree of dependency of a set of attributes Q on a set of attributes P. This can be achieved as follows: For P,$Q \subseteq A$, it can be said that Q depends on P in a degree k $(0 \leq k \leq 1)$, this is denoted $(P \Rightarrow_k Q)$ if:

$$k = \gamma_P(Q) = \frac{|POS_P(Q)|}{|U|} \quad (9)$$

The reduction of attributes can be achieved through the comparison of equivalence relations generated by sets of attributes. Attributes are removed such that the reduced set provides identical predictive capability of the decision feature or features as that of the original or unreduced set of features. A $reduct$ can be defined as a subset of minimal cardinality R_{min} of the conditional attribute set where $\gamma_R(D) = \gamma_C(D)$.

The QUICKREDUCT algorithm shown in Fig.1 searches for a minimal subset without exhaustively generating all possible subsets. The search begins with an empty subset, attributes which result in the greatest increase in the rough set dependency value are added iteratively. This process continues until the search produces its maximum possible dependency value for that dataset $(\gamma_c(\mathrm{D}))$. Note that this type of search does not guarantee a minimal subset and may only discover a local minimum.

QUICKREDUCT(C,D).
C, the set of all conditional features;
D, the set of decision features.

(1) $R \leftarrow \{\}$
(2) **do**
(3) $T \leftarrow R$
(4) $\forall x \in (C - R)$
(5) **if** $\gamma_{R \cup \{x\}}(D) > \gamma_T(D)$
(6) $T \leftarrow R \cup \{x\}$
(7) $R \leftarrow T$
(8) **until** $\gamma_R(D) == \gamma_C(D)$
(9) **return** R

Fig. 1. The QUICKREDUCT algorithm

A. Rough Set Extensions

Variable precision rough sets (VPRS) [15] attempts to introduce an element of *'fuzziness'* to the rough set model and hence (although indirectly) utilise the boundary region information. The VPRS model allows the relaxation of the subset operator, and objects can then be classified with an error smaller than a certain predefined threshold level. However, the introduction of this threshold is contrary to the rough set ideology of operating only on the information contained within the data itself. A similar approach in some respects to VPRS is the tolerance rough set model (TRSM) [12]. TRSM employs a similarity relation to minimise data as opposed to the indiscernibility relation used in classical rough-sets. This allows a relaxation in the way equivalence classes are considered.

Other hybrid approaches such as fuzzy-rough sets [6], [7], have been proposed in order to improve the ability to deal with uncertainty and vagueness present in data.

III. DISTANCE MEASURE ASSISTED ROUGH SET ATTRIBUTE REDUCTION

As discussed previously, almost all techniques for rough set attribute reduction adopt an approach to minimisation that employs the information contained within the lower approximation of a set. Very little work has been carried out where the information in the boundary region is considered for the purpose of minimisation.

The approach described below uses both the information contained in the lower approximation and the information contained in the boundary region to search for reducts. The positive region (as defined in section II) is the union of lower approximations, and this is used as described previously for the minimisation of data. The lower approximation is the set of those objects which can be said with certainty to belong to a set X. The upper approximation is the set of objects

which either definitely or possibly belong to the set X. The difference between the upper and lower approximation is the area known as the boundary region. The boundary region is an area of uncertainty.

Currently there is no mechanism in rough set based methods to deal with the uncertainty of the boundary region. Any useful information that may be contained in the boundary region is therefore lost when only the lower approximation is employed for minimisation. In order to address this, the DMRSAR method uses a distance measure to determine the proximity of objects in the boundary region to those in the lower approximation and assign a significance value to these distances.

A. Distance Metric and Mean Lower Approximation Definitions

The distance metric attempts to qualify the objects in the boundary region with regard to their proximity to the lower approximation. Similar work although that which does not specifically involve the lower approximation can be found in [13]. Intuitively, the closer the proximity of an object in the boundary region to the upper margin of the lower approximation, the higher the likelihood that it belongs to the set of interest. For the method detailed here, all of the distances of objects in the boundary region are calculated. From this the significance value for a set can be obtained.

Since calculating the margin of the lower approximation for an n-dimensional space would involve considerable computational effort, a more pragmatic solution is employed, - the mean of all object attribute values in the P-lower approximation is calculated. This can be defined as follows:

$$\underline{P}X_{MEAN} = \{\frac{\sum_{o \in \underline{P}X} a(o)}{|\underline{P}X|} : \forall a \in P\} \tag{10}$$

Using this definition of the mean of the P-lower approximation, the distance function for the proximity of objects in the boundary region from the P-lower approximation mean can be defined, $\delta_P(\underline{P}X_{MEAN}, y), \quad y \in BND_P(Q)$.

The exact function is not defined here as a number of strategies can be employed for the calculation of the distance of objects in the boundary. In the worked example section a Euclidean type distance metric is employed.

In order to measure the quality of the boundary region, a significance value ω for subset P is calculated by obtaining the sum of all object distances and inverting it such that:

$$\omega_P(Q) = (\sum_{y \in BND_P(Q))} \delta_P(\underline{P}X_{MEAN}, y))^{-1} \tag{11}$$

This significance measure takes values from the interval [0,1] and is used in conjunction with the rough set dependency value to gauge the utility of attribute subsets in a similar way to that of the rough set dependency measure. As one measure only operates on the objects in the lower approximation and the other only on the objects in the boundary, both entities are considered separately and then combined to create a new evaluation measure M:

$$M(X) = \frac{\omega_X(Q) + \gamma_X(Q)}{2} \tag{12}$$

A mean of both values is obtained as both operate in the range [0,1]. With this in mind, a new feature selection mechanism can be constructed that uses both the significance value and the rough dependency value to guide the search for the best feature subset.

B. Distance Measure-based DMQUICKREDUCT

Figure 2 below shows a rough-set based DMQUICKREDUCT algorithm based on the previously described rough algorithm in Figure 1.

DMQUICKREDUCT(C,D).
C, the set of all conditional features;
D, the set of decision features.

(1) $T \leftarrow \{\}, R \leftarrow \{\}$
(2) **do**
(3) $\quad \forall x \in (C - R)$
(4) $\quad\quad$ **if** $M(R \cup \{x\}) > M(T)$
(5) $\quad\quad\quad T \leftarrow R \cup \{x\}$
(6) $\quad R \leftarrow T$
(7) **until** $\gamma_R(D) == \gamma_C(D)$
(8) **return** R

Fig. 2. The rough-set distance metric-based QUICKREDUCT algorithm

DMQUICKREDUCT is similar to the RSAR algorithm but uses a combined distance and rough-set dependency value of a subset to guide the feature selection process. If the combined value M of the current reduct candidate is greater than that of the previous, then this subset is retained and used in the next iteration of the loop. It is important to point out that the subset is evaluated by examining the value of M, termination only occurs when the addition of any remaining features results in the dependency function value (γ_T) reaching that of the unreduced dataset. The value of M is therefore not used as a termination criterion.

The algorithm begins with an empty subset R. The do-until loop works by examining the combined dependency/significance value of a subset and incrementally adding a single conditional feature at a time. For each iteration, a conditional feature that has not already been evaluated will be temporarily added to the subset R. The combined measure of the subset currently being examined (line 6) is then evaluated and compared with that of T (the previous subset). If the combined measure of the current subset is greater, then the attribute added in (line 5) is retained as part of the new subset T (line 6).

The loop continues to evaluate in the above manner by adding conditional features, until the dependency value of the current reduct candidate ($\gamma_R(D)$) equals the consistency of the dataset (1 if the dataset is consistent).

C. A Worked Example

To illustrate the operation of the new distance measure-based algorithm, a small example dataset is considered, containing discrete-valued conditional and decision attributes. The data used in the experimentation section is real-valued, however crisp data is used in this example to aid explanation of the approach. Note also for brevity, that only the selection of two subsets is shown here.

Table I contains seven objects. It has four crisp-valued conditional attributes and a single crisp-valued decision attribute.

Object	a	b	c	d	e
0	1	0	2	2	0
1	0	1	0	0	2
2	1	0	0	1	1
3	1	0	0	2	2
4	1	2	0	0	1
5	1	2	0	2	0
6	0	1	2	0	1

TABLE I
EXAMPLE DATASET: CRISP ATTRIBUTES

If attribute d is considered for selection for example, the lower and upper approximations must first be calculated:

$\underline{\{d\}} = \{\}$ (where e = 0)
$\underline{\{d\}} = \{2\}$ (where e = 1)
$\underline{\{d\}} = \{\}$ (where e = 2)

Similarly for the upper approximation:

$\overline{\{d\}} = \{0,3,5\}$ (where e = 0)
$\overline{\{d\}} = \{1,2,4,6\}$ (where e = 1)
$\overline{\{d\}} = \{0,1,3,4,6\}$ (where e = 2)

Having calculated the upper and lower approximations for $\{d\}$, the positive and boundary regions can be shown to be:

$$POS_{\{d\}}(\{e\}) = \bigcup\{\emptyset, \{2\}\} = \{2\}$$
$$BND_{\{d\}}(\{e\}) = \bigcup\{\{0,3,5\},\{2\}\ \{1,4,6\},\{1,4,6\}\} \text{ - } \{2\} = \{0,1,3,4,5,6\}$$

The rough-set dependency, the lower approximation mean and object distances can now all be calculated. As mentioned in the previous section there are many distance metrics which can be applied to measure the distance of the objects in the boundary from the lower approximation mean. For simplicity, a variation of Euclidean distance is used in the approach documented here, and this is defined as:

$$\delta_P(\underline{P}X_{MEAN}, y) = \sqrt{\sum_{a \in P} f_a(\underline{P}X_{MEAN}, y)^2} \qquad (13)$$

where:

$$f_a(\underline{P}X_{MEAN}, y) = 1 \iff a(\underline{P}X_{MEAN}) \neq a(y)$$
$$= 0 \ otherwise$$

From this, the distances of all of the objects in the boundary region in relation to the lower approximation mean can now be calculated.

As there is only a single object in the lower approximation, the mean of the lower approximation does not need to be calculated in this case. The individual distances of objects in the boundary of $\{d\}$ can be shown to be:

$obj\ 0\ \sqrt{f_d(\underline{P}X_{MEAN}, 0)^2} = 1$
$obj\ 1\ \sqrt{f_d(\underline{P}X_{MEAN}, 1)^2} = 1$
$obj\ 3\ \sqrt{f_d(\underline{P}X_{MEAN}, 3)^2} = 1$
$obj\ 4\ \sqrt{f_d(\underline{P}X_{MEAN}, 4)^2} = 1$
$obj\ 5\ \sqrt{f_d(\underline{P}X_{MEAN}, 5)^2} = 1$
$obj\ 6\ \sqrt{f_d(\underline{P}X_{MEAN}, 6)^2} = 1$

Where there is more than one object in the potential reduct lower approximation, calculating the $\underline{P}X_{MEAN}$ object can be achieved in the manner described in the previous section. Examine all of those attribute values for each of the objects that appear in the lower approximation of the considered subset. For example considering the subset $\{a, d\}$, the lower approximation and boundary regions are:

$$POS_{\{a,d\}}(\{e\}) = \bigcup\{\emptyset, \{2\}, \{4\}\}$$
$$BND_{\{a,d\}}(\{e\}) = \bigcup\{\{0,3,5\},\{0,3,5\}\{1,6\},\{1,6\}\} = \{0,1,3,5,6\}$$

The attribute values for $\{a, d\}$ for objects $\{2, 4\}$ can be obtained by referring to Table I:

$for\ \{a\}:\ object\ 2 = '1'$
$object\ 4 = '1'$

$for\ \{d\}:\ object\ 2 = '1'$
$object\ 4 = '0'$

This results in: $\underline{P}X_{MEAN} = \{1, 0.5\}$ for $\{a, d\}$

These real-valued numbers however, are not meaningful when dealing with crisp-valued data (1 is considered as different from 1.1 as it is from 100). The strategy employed to address this problem was to examine all of the attribute values for the attribute in question and assign it a value which appears in that range of values to which it is closest in terms of magnitude. So as the $\underline{P}X_{MEAN}$ value for the attribute a is an existing value, this does not need to be considered, the $\underline{P}X_{MEAN}$ value assigned to d however is not in the range of values taken by the attribute d. Values of 0.5 or less are considered to be closer to 0, thus approximated to $'0'$, and becomes $\underline{P}X_{MEAN} = \{1, 0\}$.

Again by utilisation of Euclidean distance and the new $\underline{P}X_{MEAN}$, the distances of objects in the boundary region can be calculated:

$ob\ 0\quad \sqrt{(f_a(\underline{P}X_{MEAN}, 0)^2 + f_d(\underline{P}X_{MEAN}, 0)^2)} = 1$
$ob\ 1\quad \sqrt{(f_a(\underline{P}X_{MEAN}, 1)^2 + f_d(\underline{P}X_{MEAN}, 1)^2)} = 1$
$ob\ 3\quad \sqrt{(f_a(\underline{P}X_{MEAN}, 3)^2 + f_d(\underline{P}X_{MEAN}, 3)^2)} = 1$
$ob\ 5\quad \sqrt{(f_a(\underline{P}X_{MEAN}, 5)^2 + f_d(\underline{P}X_{MEAN}, 5)^2)} = 1$
$ob\ 6\quad \sqrt{(f_a(\underline{P}X_{MEAN}, 6)^2 + f_d(\underline{P}X_{MEAN}, 6)^2)} = 1$

It is perhaps worth noting at this point, that although a form of Euclidean distance is used to calculate the distance of the objects from the $\underline{P}X_{MEAN}$, in calculating that distance, the difference between two values is always considered in boolean terms for crisp data. The reason for this is that the values are states rather than real-valued. This means that if the value for a particular attribute in the $\underline{P}X_{MEAN}$ happened to be 1 and that of the corresponding attribute value of an object in the boundary region was 1563, the difference between these two states is $(1 - 1563) = 1$.

Although the individual distances may be useful in identifying objects that are similar to those in the lower approximation, they are not individually indicative of the subset goodness. A method of achieving this measure is to calculate the sum of all of the distances and invert it thus giving a significance value to each subset considered for selection. The significance value is real-valued and has membership in the range [0,1] for the purpose of dealing with crisp data.

Thus for $\{a, d\}$:

$$\omega_{\{a,d\}}(\{e\}) \quad = \quad (\sum(1, 1, 1, 1, 1))^{-1} \quad = 0.2$$

Although the significance measure alone can be used to search for subsets, empirical results demonstrated that these were not of equal quality as those returned by RSAR. So the significance value was combined with the rough set dependency value. This results in a combined metric in which both dependency and significance have equal participation.

By calculating the change in combined significance and dependency value (M) when an attribute is removed from the set of considered conditional attributes, a measure of the goodness of that attribute can be obtained. The greater the change in M the greater the measure of goodness that attribute has attached to it.

Using the previous examples of the DMRSAR method the values for the combined metric can be calculated for all considered subsets of C using DMRSAR:

$$\begin{array}{rr} M_{\{b\}}(\{e\}) = 0.0 & M_{\{b,d\}}(\{e\}) = 0.3910 \\ M_{\{c\}}(\{e\}) = 0.0 & M_{\{c,d\}}(\{e\}) = 0.3026 \\ M_{\{d\}}(\{e\}) = 0.342 & M_{\}a,b,d\}}(\{e\}) = 0.3026 \\ M_{\{a,d\}}(\{e\}) = 0.2425 & M_{\{b,c,d\}}(\{e\}) = 1.0 \end{array}$$

It is obvious from the above example that the search finds a subset in the manner $\{d\} \rightarrow \{b, d\} \rightarrow \{b, c, d\}$. As $\{a, d\}$ and $\{c, d\}$ and also $\{a, b, d\}$ do not result in the same increase in combined metric these subsets are ignored.

IV. Experimentation

This section presents the results of experimental studies using the 8 real-valued datasets. It is important to note that DMRSAR operates on discretised versions of the datasets listed. These datasets are of the same format as that used in the example in the previous section. They are small-to-medium in size, with between 120 and 390 objects per dataset and feature sets ranging from 5 to 39. All datasets have been obtained from [1] and [8]. A comparison of the FRFS algorithm and the distance-based dimensionality reduction techniques is given based on subset size, classification accuracy, and time taken to discover subsets.

Dataset	Unreduced	QSBA FRFS	DMRSAR
water 2	57.940	61.282	61.282
water 3	48.971	38.710	43.330
cleveland	37.459	39.908	33.850
glass	43.650	37.010	39.696
heart	64.070	67.037	65.799
iris	80.670	86.000	78.667
olitos	64.166	59.106	50.833
wine	94.860	88.202	85.490

TABLE II
AVERAGE CLASSIFICATION ACCURACY

A. Classifier

In the generation of results for classification accuracies, a single fuzzy classifier QSBA [10] was used, as this was readily available, although other fuzzy classification systems could be employed for this purpose.

QSBA works by generating fuzzy rules using the fuzzy subsethood measure for each decision class and a threshold to determine what appears in the rule for that decision class. The fuzzy subsethood measure is then used to act as weights, and the algorithm then modifies the weights to act as fuzzy quantifiers.

B. Comparison of Classification Accuracy

The data presented in Table II shows the average classification accuracy as a percentage obtained using the 10-fold cross validation method. The classification was initially performed on the unreduced dataset, followed by the reduced datasets which were obtained, by using both the FRFS and DMRSAR dimensionality reduction techniques respectively.

It is interesting to note that where a decrease in classification accuracy is recorded for FRFS, with respect to the unreduced data the same is also true for DMRSAR. This fall in classification accuracy is small when comparing both FRFS and DMRSAR approaches to the unreduced data. Also when comparing classification results, where the DMRSAR approach shows a fall in classification accuracy, the corresponding reduction in dimensionality (shown in Table III) is significantly better than that of FRFS.

C. Subset size, and Run times

Presented in Table III is a comparison of subset size, and runtime data, for both FRFS and DMRSAR approaches. There is an obvious and clear advantage to the DMRSAR approach in relation to subset size. There are two datasets where this is not the case water 2 and water 3. These datasets are difficult to manipulate, however there has been no effort to optimise the DMRSAR approach and it is expected that gains could be made in this respect through the implementation of such improvements.

It is clear also from the runtime figures that DMRSAR runs considerably faster than FRFS, with all but water-2 running in sub 1-second times. This primarily, can be attributed to

Dataset	Original number of features	objects	Subset size FRFS	DMRSAR	Time taken to locate subset FRFS	DMRSAR
water 2	39	390	11	12	96.58	0.860
water 3	39	390	12	23	158.73	1.266
cleveland	14	297	11	9	24.11	0.219
glass	10	214	9	6	1.61	0.156
heart	14	270	11	10	11.84	0.158
iris	5	150	5	4	0.031	0.062
olitos	26	120	10	8	11.20	0.156
wine	14	178	10	8	1.42	0.125

TABLE III
COMPARISON OF SUBSET SIZE, DEPENDENCY VALUE, & RUN TIMES

the computational complexity of FRFS. Considering also that no runtime optimisation has been performed for DMRSAR these results are very encouraging.

V. CONCLUSIONS

Comparison of both FRFS and DMRSAR has shown that the DMRSAR method is a good starting point for further work based on the distance metric for investigating the quality of reducts. The subset size results show that there is still some additional optimisation required in order to equal FRFS.Classification accuracy results have been shown to be very similar to those of FRFS, and in some cases the DMRSAR method has even shown an increase whilst simultaneously demonstrating a reduction in dimensionality. Where a decrease has been observed in relation to FRFS, it has been small and, as discussed previously, the actual decrease is not significant.

It is clear from the results obtained in the previous section that an increase in the efficiency of the DMRSAR algorithm is highly desirable. The experimental work detailed in this paper did not take advantage of any optimisations that are expected would improve the performance of DMRSAR further.

Future work would include a re-evaluation of how the mean lower approximation, is calculated. Implementation of a more accurate calculation of the lower approximation boundary would mean that distances of objects in the boundary region could be more accurately measured.

The significance measure which is employed for DMRSAR is also very basic, and considers the boundary region as a single significance value which is expressed as membership value of a unary fuzzy set. By redefining this as a number of fuzzy sets, the boundary region could be quantified more accurately by expressing membership in terms of weights of objects in the boundary in relation to distance from the lower approximation. Apart from the use of extra fuzzy sets, the way in which objects in the boundary are related is another area which is worthy of investigation. By examining the correlation of objects and their individual distances, it may be possible to qualify the individual objects and their information value.

Other areas worthy of investigation include the distance metric and also the application area of the approach. For the worked example described in this paper a Euclidean distance metric is employed. Metrics such as Mahalanobis distance, ellipsoid distance, and others could also be considered. Additionally, the distance-based rough set approach, is equally applicable to areas such as clustering.

REFERENCES

[1] C. Armanino, R. Leardi, S. Lanteri, and, G. Modi Chemom.Intell. Lab.Syst. , vol. 5, pp. 343–354. 1989.

[2] A. Chouchoulas and Q. Shen. Rough set-aided keyword reduction for text categorisation. Applied Artificial Intelligence, Vol. 15, No. 9, pp. 843–873. 2001.

[3] J.S. Deogun, V.V. Raghavan, and H. Sever. Exploiting Upper Approximation in the Rough Set Methodology. Proceedings of the First International Conference on Knowledge Discovery and Data Mining, (Montreal, Quebec, Canada), pp. 1–10, 1995

[4] P. Devijver and J. Kittler. Pattern Recognition: A Statistical Approach. Prentice Hall. 1982.

[5] M. Inuiguchi and M. Tsurumi. Measures Based on Upper Approximations of Rough Sets for Analysis of Attribute Importance and Interaction International Journal of Innovative Computing, Information and Control, Vol. 2, No. 1, pp 1–12. 2006

[6] R. Jensen and Q. Shen. Fuzzy-Rough Attribute Reduction with Application to Web Categorization. Fuzzy Sets and Systems, Vol. 141, No. 3, pp. 469–485. 2004.

[7] N. Mac Parthaláin, R. Jensen, and Q. Shen. Fuzzy Entropy-Assisted Fuzzy-Rough Feature Selection. Proceedings of the 15th International Conference on Fuzzy Systems (FUZZ-IEEE'06). 2006.

[8] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz, UCI Repository of machine learning databases [http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science. 1998.

[9] Z. Pawlak, Rough sets. Int. J. Comput. Inf. Sci. 11, pp. 341–356. 1982.

[10] K.A. Rasmani and Q. Shen. Data-driven fuzzy rule generation and its application for student academic performance evaluation. Applied Intelligence, vol. 25, no. 3, pp. 305–319, 2006.

[11] Q. Shen and R. Jensen. Selecting Informative Features with Fuzzy-Rough Sets and its Application for Complex Systems Monitoring. Pattern Recognition, Vol. 37, No. 7, pp. 1351–1363. 2004.

[12] A. Skowron, J. Stepaniuk, Tolerance Approximation Spaces, Fundamenta Informaticae, Vol. 27, pp. 245-253, 1996.

[13] D. Slezak, Various Approaches to Reasoning with Frequency Based Decision Reducts: A Survey, in: L. Polkowski, S. Tsumoto, T.Y. Lin, (eds.), Rough Set Methods and Applications, Heidelberg: Physica-Verlag, pp 235–285. 2000.

[14] I.H. Witten and E. Frank. Generating Accurate Rule Sets Without Global Optimization. In Machine Learning: Proceedings of the 15th International Conference, Morgan Kaufmann Publishers, San Francisco. 1998.

[15] W. Ziarko. Variable Precision Rough Set Model. Journal of Computer and System Sciences, Vol. 46, No. 1, pp. 3959. 1993.